Variational Bayesian Image Restoration Based on a Product of t-Distributions Image Prior

Giannis Chantas, Nikolaos Galatsanos, Aristidis Likas, and Michael Saunders

***Abstract*—Image priors based on products have been recognized to offer many advantages because they allow simultaneous enforcement of multiple constraints. However, they are inconvenient for Bayesian inference because it is hard to find their normalization constant in closed form. In this paper, a new Bayesian algorithm is proposed for the image restoration problem that bypasses this difficulty. An image prior is defined by imposing Student-t densities on the outputs of local convolutional filters. A variational methodology, with a constrained expectation step, is used to infer the restored image. Numerical experiments are shown that compare this methodology to previous ones and demonstrate its advantages.**

***Index Terms*—Constrained variational inference, image restoration, product prior, Student's-t prior, Variational Bayesian Inference.**

I. Introduction

IMAGE restoration is a well known ill-posed inverse problem that requires regularization. Regularization based on Bayesian methodology is very popular since it provides a systematic and rigorous framework for estimation of the model parameters. Regularization in a Bayesian framework corresponds to the introduction of a prior for the image statistics [1], which enforces prior knowledge for the image.

Initially, stationary Gaussian priors were used; see for example [2] and [3]. Such priors are convenient from an implementation point of view because they require only one parameter; however, they have the drawback of not being able to preserve edges and they smooth noise in flat areas of the image. To avoid this problem, there has been a very large body of work in the last 20 years. A number of methods have been introduced to regularize in a spatially variant manner, or equivalently, many edge-preserving priors have been proposed. A detailed survey on this topic is beyond the scope of this paper. In what follows, we selectively reference work that is pertinent to the herein proposed approach.

Manuscript received January 7, 2008; revised June 7, 2008. Current version published September 10, 2008. This work was supported in part by the E.U.-European Social Fund (75%) and in part by the Greek Ministry of Development-GSRT (25%). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Elad.

G. Chantas and A. Likas are with the Department of Computer Science, University of Ioannina, Ioannina, Greece, 45110 (e-mail: chanjohn@cs.uoi.gr; arly@cs.uoi.gr).

N. Galatsanos is with the Department of Electrical and Computer Engineering, University of Patras, Rio, 26500, Greece (e-mail: ngalatsanos@upatras.gr).

M. Saunders is with the Department of Management Science and Engineering (MS&E), Stanford University, Stanford, CA USA 94305-4026 (e-mail: saunders@stanford.edu).

Digital Object Identifier 10.1109/TIP.2008.2002828

Priors based on robust Huberov statistics and Generalized Gaussian pdfs have also been used; see, for example, [4] and [5]. Recently, such a prior was used along with the majorization minimization framework to derive an edge-preserving image restoration algorithm that can be implemented very efficiently using the fast Fourier transform [6]. The main shortcomings of such priors are that their normalization constant is hard to find. The parameters of such models have to be adjusted empirically.

Another class of algorithms, which have been very popular in certain image processing circles designed for edge-preserving restoration, is based on the total variation (TV) criterion [8]. Although TV-based regularization has been popular, till recently it involved ad hoc selection of certain parameters. Recently, though, a Bayesian framework was proposed that allows estimation of these parameters in a rigorous manner. Nevertheless, improper priors were used in these works, and as a result these methodologies contain an element of subjective selection [9], [10] and [11].

Priors based on wavelet decompositions and heavy-tailed pdfs have been used for edge-preserving image restoration in [12] and [13] along with the EM algorithm. In [7] and [14], the denoising problem was addressed with heavy-tailed priors in the wavelet domain. Image denoising involves a simpler imaging model; as a result Bayesian inference is easier in this case. A Gaussian scale mixture (GSM) was used to model the wavelet coefficients in [7] and a one-step algorithm for inference. In [14], Hirakawa and Meng used the Student-t pdf to model the statistics of the wavelet coefficients, and derived an EM algorithm for inference. The Student's-t pdf is a special case of a GSM.

Product-based image priors have also been proposed in [15]. Such priors combine in product form multiple probabilistic models. Each individual model gives high probability to data vectors that satisfy just one constraint. Vectors that satisfy only this constraint but violate others are ruled out by their low probability under the other terms of the product model. Image priors based on this idea have been used in image recovery problems [15] and [16]. However, such priors were learned using a large training set with images and stochastic sampling methods and used in a number of image recovery problems based on "empirical" maximum *a posteriori* approaches and gradient descent minimization [15]. This differs from the herein proposed approach where the product prior is learnt only from the observations. The term "empirical" is used because the PoE priors used were not normalized; thus, the parameters of the recovery algorithm cannot be estimated or inferred rigorously but were adjusted rather empirically.

In [17], [18], and [19], some of us proposed a new hierarchical image prior for image restoration, image super-resolution, and

blind image deconvolution problems, respectively. This prior is Student-t based, is in a product form, and is able to capture the local image discontinuities and thus provide edge-preserving capabilities for those problems. Its main shortcomings are that both the normalization constant and the hyper-parameters of the prior were found heuristically. Furthermore, image models based on Student-t statistics have been used with success in other than image reconstruction applications. For example, in [21], such models were used with success for watermark detection.

Inspired by our previous work, we now propose a new Bayesian inference framework for image deconvolution using a prior in product form. This prior assumes that the outputs of local high-pass filters are Student-t distributed. The main contribution of this work is a Bayesian inference methodology that bypasses the difficulty of evaluating the normalization constant of product type priors. The methodology is based on a *constrained* variational approximation that uses the outputs of all the local high pass filters to produce an estimate of the original image. More specifically, a *constrained expectation step* is used to capture the relationship of the filter outputs of the prior to the original image. In this manner, the use of improper priors is avoided and *all* the parameters of the prior model are estimated from the data. Thus, the "trial and error" parameter "tweaking" required in [17]–[19] and other state-of-the-art recently proposed restoration algorithms, which makes their use difficult use for nonexperts, is avoided. Furthermore, the proposed restoration algorithm provides competitive performance compared with previous methods.

In this work, we also propose an efficient Lanczos-based computational framework tailored to the calculations required in our Bayesian algorithm. More specifically, a very large linear system $Ax = b$ is solved iteratively and the diagonal elements of a matrix $Q^t A^{-1} Q$ are simultaneously estimated in an efficient manner.

The rest of this paper is organized as follows. In Section II, the imaging and image model are defined. In Section III, the variational restoration algorithm is derived. In Section IV, we present the computational methodology used to implement our algorithm. In Section V, numerical experiments are demonstrated. Finally, Section VI gives conclusions and thoughts for future work.

II. Imaging and Image Model

A. Imaging Model

A linear imaging model is assumed. For convenience but without loss of generality, we use 1-D notation. The $N \times 1$ vector $\mathbf{g}$ represents the observed degraded image obtained by

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \mathbf{n} \tag{2.1}$$

where $\mathbf{f}$ is the (unknown) original image, $\mathbf{H}$ is an $N \times N$ known convolution matrix and $\mathbf{n}$ is additive white noise. We assume Gaussian statistics for the noise given by $\mathbf{n} \sim N(\mathbf{0}, \beta^{-1}\mathbf{I})$ where $\mathbf{0}$ is an $N \times 1$ vector of zeros, $\mathbf{I}$ is the $N \times N$ identity matrix and β is the noise precision (inverse variance), which is assumed unknown.

Aiming at the definition of the image prior we first define P operators $\mathbf{Q}_k$ for $k = 1, \ldots, P$ and use them to define P filter outputs

$$\boldsymbol{\varepsilon}_k = \mathbf{Q}_k \mathbf{f}, \quad k = 1, \ldots, P \tag{2.2}$$

where $\boldsymbol{\varepsilon}_k = [\varepsilon_k(1), \varepsilon_k(2), \ldots, \varepsilon_k(N)]^T$. The matrices $\mathbf{Q}_k$ representing the operators are of size $N \times N$ and the filter outputs $\boldsymbol{\varepsilon}_k$ are of size $N \times 1$. These operators are zero mean convolutional high-pass filters and each one of them is used to impose a particular constraint on the restored image.

B. Image Prior Model

We assume that $\varepsilon_k(i)$ for $i = 1, \ldots, N$ are i.i.d zero mean Student-t distributed, with parameters λ_k and ν_k

$$\varepsilon_k(i) \sim \mathrm{St}(\varepsilon_k(i); 0, \lambda_k, \nu_k), \quad \forall i, \quad \forall k \tag{2.3}$$

where

$$\mathrm{St}(x; 0, \lambda_k, \nu_k) = \frac{\Gamma(\nu_k/2 + 1/2)}{\Gamma(\nu_k/2)} \times \left(\frac{\lambda_k}{\pi\nu_k}\right)^{1/2} \left(1 + \frac{\lambda_k}{\nu_k} x^2\right)^{-\frac{\nu_k+1}{2}}.$$

The Student-t implies a two-level generation process [22]. More specifically, $a_k(i)$ is first drawn from a Gamma distribution, $p(a_k(i)) = \mathrm{Gamma}(a_k(i); \nu_k/2, \nu_k/2)$. Then, the $\varepsilon_k(i)$ is generated from a zero-mean Normal distribution with precision $\lambda_k a_k(i)$, according to $p(\varepsilon_k(i)|a_k(i)) = N(\varepsilon_k(i); 0, (\lambda_k a_k(i))^{-1})$. The probability density function of (2.3) can be written as an integral

$$p(\varepsilon_k(i)) = \mathrm{St}(\varepsilon_k(i); 0, \lambda_k, \nu_k) = \int_0^{+\infty} p(\varepsilon_k(i)|a_k(i)) p(a_k(i))\, da_k(i).$$

The variables $a_k(i)$ are called "hidden" (latent) because they are not apparent in (2.3), since they have been integrated out. There are two extremes in this generative model, depending on the value of the "degree of freedom" parameter ν_k. As this parameter goes to infinity, the pdf from which the $a_k(i)$'s are drawn has its mass concentrated around 1. This in turn reduces the Student-t to a Normal distribution, because all $\varepsilon_k(i)$ are drawn from the same Normal with precision λ_k, since $a_k(i) = 1$. The other extreme is when $\nu_k \to 0$ and the prior becomes uninformative. In general, for small values of ν_k the probability mass of the Student-t pdf is spread, rendering the Student-t more "heavy-tailed".

The use of heavy-tailed priors on high-pass filters of the image is a characteristic of most modern "edge preserving" image priors used for regularization in a stochastic setting; see for example [4]–[6], [11], [14], [15], and [19]. The main idea behind this assumption is that at the few edge areas of an image the filter outputs $\varepsilon_k(i)$ will be large in absolute value. Thus, it is important to model them with a heavy-tailed pdf in order to allow the prior to encourage formation of edges. The downside

of many such models is that most heavy-tailed pdfs are not amenable to Bayesian inference. For example, the Generalized Gaussian and the Alpha Stable pdfs can be also heavy tailed. However, unlike the Student-t where Bayesian inference is possible [27], moment-based estimators have to be used for their parameters; see for example [24] and [25].

We now define the following notation for the variables $a_k(i)$. We denote by $\tilde{\mathbf{a}} = [\mathbf{a}_1, \ldots, \mathbf{a}_P]^{\mathrm{T}}$ a $PN \times 1$ vector, where $\mathbf{a}_k = [a_k(1), a_k(2), \ldots, a_k(N)]$. Also, for the filter outputs we use the notation $\tilde{\boldsymbol{\varepsilon}} = [(\boldsymbol{\varepsilon}_1)^{\mathrm{T}}, (\boldsymbol{\varepsilon}_2)^{\mathrm{T}}, \ldots, (\boldsymbol{\varepsilon}_P)^{\mathrm{T}}]^{\mathrm{T}}$. We assume that the filter outputs are independent not only in each pixel location but also in each direction. This assumption makes subsequent calculations tractable. Thus, the cumulative density for the filter outputs conditioned on $\tilde{\mathbf{a}}$ is

$$p(\tilde{\boldsymbol{\varepsilon}}|\tilde{\mathbf{a}}) = \prod_{k=1}^{P} p(\boldsymbol{\varepsilon}_k|\mathbf{a}_k) \tag{2.4}$$

where $p(\boldsymbol{\varepsilon}_k|\mathbf{a}_k) = N(\mathbf{0}, (\lambda_k \mathbf{A}_k)^{-1})$ and $\mathbf{A}_k$ is a diagonal matrix with elements the components of the vector $\mathbf{a}_k$.

At this point, the marginal distribution $p(\mathbf{f})$ yearns for a closed form, using the relation between the image and the filter outputs, (2.2). However, this prior is analytically intractable because *one cannot find in closed form its normalization constant*. This problem stems from the fact that it is not possible to find the eigenvalues of the matrix $\sum_{k=1}^{P} \mathbf{Q}_k^{\mathrm{T}} \mathbf{A}_k \mathbf{Q}_k$ because it is very large and the product $\mathbf{Q}_k^{\mathrm{T}} \mathbf{A}_k \mathbf{Q}_k$ does not have a structure that is amenable to efficient eigenvalue computation. One contribution of this work is that we bypass this difficulty by exploiting the commuting property of convolutional operators and derive a constrained variational algorithm for approximate Bayesian inference. This algorithm is described in detail next.

III. Variational Algorithm

Since, as explained above, it is difficult to infer a solution for the image from the Bayesian model previously defined, a transformed imaging model is introduced in Section 3.1.

A. Variational Algorithm for the Equivalent Imaging Model

The imaging model of (2.1) can be written as

$$\mathbf{Q}_k \mathbf{g} = \mathbf{Q}_k \mathbf{H} \mathbf{f} + \mathbf{Q}_k \mathbf{n} \quad \text{for } k = 1, \ldots, P. \tag{3.1}$$

Setting $\mathbf{y}_k = \mathbf{Q}_k \mathbf{g}$ for $k = 1, \ldots, P$ and using (2.2), we can utilize the *commuting property* of the convolutional operators and write the imaging model as

$$\mathbf{y}_k = \mathbf{H} \boldsymbol{\varepsilon}_k + \mathbf{n}_k \quad \text{for } k = 1, \ldots, P \tag{3.2}$$

where $\mathbf{y}_k$ are the observations of the newly defined model and the additive noise is

$$\mathbf{n}_k \sim N\left(0, \beta^{-1} \mathbf{Q}_k \mathbf{Q}_k^{\mathrm{T}}\right).$$

In this model, we assume that the filter outputs $\boldsymbol{\varepsilon}_k$ of our filters $\mathbf{Q}_k$ are the unknowns. Thus, the algorithm will infer $\tilde{\boldsymbol{\varepsilon}}$ instead of $\mathbf{f}$. In this manner we bypass the need to define a prior for $\mathbf{f}$. For this reason, we must initially define the posterior of the observations $\tilde{\mathbf{y}}$ given $\tilde{\boldsymbol{\varepsilon}}$. This is equal to the product of P Normal distributions, since the observations are assumed indepedent $\mathbf{n}$:

$$p(\tilde{\mathbf{y}}|\tilde{\boldsymbol{\varepsilon}}) = \prod_{k=1}^{P} p(\mathbf{y}_k|\boldsymbol{\varepsilon}_k), \text{ where}$$

$$\tilde{\mathbf{y}} = \left[(\mathbf{y}_1)^{\mathrm{T}}, (\mathbf{y}_2)^{\mathrm{T}}, \ldots, (\mathbf{y}_P)^{\mathrm{T}}\right]^{\mathrm{T}} \text{ and}$$

$$p(\mathbf{y}_k|\boldsymbol{\varepsilon}_k) = N\left(\mathbf{H}\boldsymbol{\varepsilon}_k, \left(\beta \mathbf{Q}_k \mathbf{Q}_k^{\mathrm{T}}\right)^{-1}\right) \quad \text{for } k = 1, \ldots, P.$$

The prior for the residuals has been already defined in (2.3).

Working in the Bayesian framework, we define as latent (hidden) variables the residuals $\tilde{\boldsymbol{\varepsilon}}$ and the inverse variances $\tilde{\mathbf{a}}$. Hence, the complete data likelihood is

$$p(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}; \theta) = p(\tilde{\mathbf{y}}|\tilde{\boldsymbol{\varepsilon}}; \theta) p(\tilde{\boldsymbol{\varepsilon}}|\tilde{\mathbf{a}}; \theta) p(\tilde{\mathbf{a}}; \theta)$$

where $\theta = [\beta, \nu_1, \ldots, \nu_P, \lambda_1, \ldots, \lambda_P]^{\mathrm{T}}$.

Estimation of the model parameters ideally could be obtained through maximization of the marginal distribution of the observations $p(\tilde{\mathbf{y}}; \theta)$

$$\hat{\theta} = \arg\max_{\theta} \int \int p(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}; \theta)\, d\tilde{\boldsymbol{\varepsilon}}\, d\tilde{\mathbf{a}}. \tag{3.3}$$

However, in the present case, this marginalization is not possible. Furthermore, since the posterior of the hidden variables given the observations $p(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}|\tilde{\mathbf{y}})$ is not known explicitly, inference via the Expectation-Maximization (EM) algorithm is not possible [29].

For this reason, we resort to the variational methodology [22], [28] and [29]. According to this methodology, we introduce a lower bound on the logarithm of the marginal likelihood, which is actually the expectation of the logarithm of the complete data likelihood with respect to an auxiliary function of the hidden variables $q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}})$ minus the entropy of $q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}})$

$$\begin{aligned} \log p(\tilde{\mathbf{y}}; \theta) &\geq L(q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}), \theta) L(q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}), \theta) \\ &\equiv \int q(\bar{\boldsymbol{\varepsilon}}, \bar{\mathbf{a}}) \log p(\bar{\mathbf{y}}, \bar{\boldsymbol{\varepsilon}}, \bar{\mathbf{a}}; \theta) d\bar{\boldsymbol{\varepsilon}} d\bar{\mathbf{a}} \\ &\quad - \int q(\bar{\boldsymbol{\varepsilon}}, \bar{\mathbf{a}}) \log q(\bar{\boldsymbol{\varepsilon}}, \bar{\mathbf{a}}) d\bar{\boldsymbol{\varepsilon}} d\bar{\mathbf{a}}. \end{aligned} \tag{3.4}$$

The inequality holds because the functional L is also equal to the logarithm of the marginal likelihood minus the always nonnegative Kullback–Leibler divergence between the true posterior distribution $p(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}|\tilde{\mathbf{y}}; \theta)$ of the hidden variables and $q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}})$; see for example [22].

Equality holds in (3.4) when $q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}) = p(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}|\tilde{\mathbf{y}}; \theta)$, or equivalently

$$q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}) = \prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k, \mathbf{a}_k) = \prod_{k=1}^{P} p(\boldsymbol{\varepsilon}_k, \mathbf{a}_k|\mathbf{y}_k; \theta_k) \tag{3.5}$$

because in this case the Kullback–Leibler divergence becomes zero.

In the variational Bayesian framework, instead of maximizing the unobtainable marginal likelihood, we maximize the bound L, (3.4), with respect to both $q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}})$ and θ in the variational E and M steps, respectively. In other words, the unknown posterior $p(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}|\tilde{\mathbf{y}}; \theta)$ is approximated by $q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}})$. One difficulty in this approach is that the maximization with respect to $q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}})$ is hard to obtain in closed form, although we can bypass it by using the so-called Mean Field approximation [29]. According to this approximation, if we assume that

$$q(\boldsymbol{\varepsilon}_k, \mathbf{a}_k) = q(\boldsymbol{\varepsilon}_k) q(\mathbf{a}_k), \text{ for } k = 1, \ldots, P \tag{3.6}$$

then unconstrained optimization of the functional $L(q(\tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}), \theta)$ with respect to all $q(\boldsymbol{\varepsilon}_k)$ yields P Normal distributions

$$q(\boldsymbol{\varepsilon}_k) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \text{for } k = 1, \ldots, P \tag{3.7}$$

with parameters $\boldsymbol{\mu}_k = \beta \boldsymbol{\Sigma}_k \mathbf{H}^{\mathrm{T}} \mathbf{g}$ and $\boldsymbol{\Sigma}_k = (\beta \mathbf{H}^{\mathrm{T}} \mathbf{H} + \lambda_k \mathbf{Q}_k^{\mathrm{T}} \mathbf{A}_k \mathbf{Q}_k)^{-1}$.

The difficulty that we encounter with the above posteriors, which were obtained by unconstrained optimization, is that they do not provide a method to infer $\mathbf{f}$ from $\boldsymbol{\varepsilon}_k$, and they do not capture their common origin from $\mathbf{f}$, (2.2).

In order to bypass this difficulty we make the assumption that each of the posteriors $q(\boldsymbol{\varepsilon}_k)$ is Normal; however, it is *constrained* so that it captures the common origin of all $\boldsymbol{\varepsilon}_k$ from $\mathbf{f}$, as dictated by (2.2). In other words, we assume that

$$q(\boldsymbol{\varepsilon}_k; \mathbf{m}, \mathbf{R}) = N\left(\mathbf{Q}_k \mathbf{m}, \mathbf{Q}_k \mathbf{R} \mathbf{Q}_k^{\mathrm{T}}\right) \quad \text{for } k = 1, \ldots, P, \tag{3.8}$$

where $\mathbf{m}$ and $\mathbf{R}$ are actually parameters representing the mean and covariance of the image $\mathbf{f}$, from which all $\boldsymbol{\varepsilon}_k$ originate. In other words

$$E[\boldsymbol{\varepsilon}_k] = \mathbf{Q}_k E[\mathbf{f}] = \mathbf{Q}_k \mathbf{m},$$
$$\mathrm{Cov}[\boldsymbol{\varepsilon}_k] = \mathbf{Q}_k \mathrm{Cov}[\mathbf{f}] \mathbf{Q}_k^{\mathrm{T}} = \mathbf{Q}_k \mathbf{R} \mathbf{Q}_k^{\mathrm{T}}.$$

Thus, $\mathbf{m}$ and $\mathbf{R}$ are parameters that are used in our model and estimated during the restoration algorithm. Actually, the restored image is taken to be the estimate of $\mathbf{m}$.

B. Variational Update Equations

The general variational algorithm using the Mean Field approximation [29] for approximate inference of a statistical model with $\mathbf{y}$ as observation, n hidden variables $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and parameters denoted by θ, aims to maximize the bound

$$L(q(\mathbf{x}), \theta) \equiv \int \prod_{i=1}^{n} q_i(\mathbf{x}) \log p(\mathbf{y}, \mathbf{x}; \theta) d\mathbf{x}_i - \int \prod_{i=1}^{n} q_i(\mathbf{x}_i) \log q(\mathbf{x}_i)\, d\mathbf{x}_i.$$

This is achieved by iterating between the two following steps, where (t) is the iteration index:

$$\mathrm{VE{-}step}: q_i^{(t+1)}(\mathbf{x}) = \arg\max_{q_i(\mathbf{x})} L\left(q(\mathbf{x}), \theta^{(t)}\right), i = 1, \ldots, n,$$
$$\mathrm{VM{-}step}: \theta^{(t+1)}(\mathbf{x}) = \arg\max_{\theta} L\left(q^{(t+1)}(\mathbf{x}), \theta\right).$$

Thus, in the E-step of the variational algorithm, optimization of the functional is performed with respect to the auxiliary functions. However, in the present case, the functions $q(\boldsymbol{\varepsilon}_k), k = 1, \ldots, P$, are assumed to be Normal distributions with partially common mean and covariance [see (3.8)]; therefore, this bound is actually a function of the parameters $\mathbf{R}$ and $\mathbf{m}$ and a functional w.r.t. the auxiliary function $q(\tilde{\mathbf{a}})$. Using (3.6), the variational bound in our problem becomes

$$L(q(\tilde{\mathbf{a}}), \theta_1, \theta_2) = \int \prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k; \theta_1) q(\mathbf{a}_k) \log p(\check{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}; \theta_2) d\tilde{\boldsymbol{\varepsilon}} d\tilde{\mathbf{a}} - \int \prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k; \theta_1) q(\mathbf{a}_k) \log \prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k; \theta_1) q(\mathbf{a}_k) d\tilde{\boldsymbol{\varepsilon}} d\tilde{\mathbf{a}} \tag{3.9}$$

where $\theta_1 = [\mathbf{R}, \mathbf{m}]$ and $\theta_2 = [\beta, \lambda_1, \ldots, \lambda_P, \nu_1, \ldots, \nu_P]^{\mathrm{T}}$. Thus, in the VE-step of our algorithm the bound must be optimized with respect to $\mathbf{R}, \mathbf{m}$, and $q(\tilde{\mathbf{a}})$.

$$\mathrm{VE{-}step}: \left[q^{(t+1)}(\tilde{\mathbf{a}}), \theta_1^{(t+1)}\right] = \arg\max_{[q(\tilde{\mathbf{a}}), \theta_1]} L\left(q(\tilde{\mathbf{a}}), \theta_1, \theta_2^{(t)}\right).$$

Taking the derivative of L w.r.t to $\mathbf{m}, \mathbf{R}$ and $q(\tilde{\mathbf{a}})$ (see Appendix), we find that the bound is maximized w.r.t. these parameters when

$$\mathbf{m}^{(t+1)} = \beta^{(t)} \mathbf{R}^{(t)} \mathbf{H}^{\mathrm{T}} \mathbf{g}, \tag{3.10}$$

$$\mathbf{R}^{(t+1)} = \left(\beta^{(t)} \mathbf{H}^{\mathrm{T}} \mathbf{H} + \frac{1}{P} \sum_{k=1}^{P} \lambda_k \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k^{(t)} \mathbf{Q}_k\right)^{-1} \tag{3.11}$$

$$q^{(t+1)}(a_k(i)) = \mathrm{Gamma}\left(a_k(i); \frac{\nu_k^t}{2} + \frac{1}{2}, \frac{\nu_k^t}{2} + \frac{1}{2} \lambda_k^t \times \left(\left(\mathbf{m}_k^{(t)}(i)\right)^2 + \mathbf{C}_k^{(t)}(i, i)\right)\right) \quad \forall k, \forall i \tag{3.12}$$

where $\mathbf{m}_k^{(t)} = \mathbf{Q}_k \mathbf{m}^{(t)}$ and $\mathbf{C}_k^{(t)} = \mathbf{Q}_k \mathbf{R}^{(t)} \mathbf{Q}_k^{\mathrm{T}}$. Notice that since each $q^{t+1}(a_k(i))$ is a Gamma pdf of the form

$q^{t+1}(a_k^{(t+1)}(i)) = \text{Gamma}(a_k^{(t+1)}(i); \alpha, \beta)$, its expected value is

$$\langle a_k(i)\rangle_{q^{t+1}(a_k(i))} = \frac{\alpha}{\beta} = (\nu_k^t + 1) \times \left(\nu_k^t + \lambda_k^t \left(\left(\mathbf{m}_k^{(t)}(i)\right)^2 + \mathbf{C}_k^{(t)}(i,i)\right)\right)^{-1} \quad (3.13)$$

where $\langle\cdot\rangle_{q(\cdot)}$ denotes the expectation w.r.t. an arbitrary distribution $q(\cdot)$. This is used in (3.10) and (3.11), where $\hat{\mathbf{A}}_k^{(t)}$ is a diagonal matrix with elements

$$\hat{\mathbf{A}}_k^{(t)}(i,i) = \langle a_k(i)\rangle_{q^{(t)}(a_k(i))}, i = 1, \ldots, N.$$

At the variational M-step the bound is maximized with respect to the model parameters

$$\text{VM}-\text{step}: \theta_2^{(t+1)} = \arg\max_{\theta_2} L\left(q^{(t+1)}(\tilde{\mathbf{a}}), \theta_1^{(t+1)}, \theta_2\right)$$

where $L(q^{(t+1)}(\tilde{\mathbf{a}}), \theta_1^{(t+1)}, \theta_2) \propto \langle \log p(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}; \theta_2)\rangle_{q(\tilde{\boldsymbol{\varepsilon}}; \theta_1^{(t+1)}), q^{(t+1)}(\tilde{\mathbf{a}})}$ is calculated using the results from (3.10)–(3.13).

The update for β is obtained after taking the derivative and equating to zero

$$\beta^{(t+1)} = N\left(\left\|\mathbf{H}\mathbf{m}^{(t+1)} - \mathbf{g}\right\|_2^2 + \text{trace}\left\{\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{R}^{(t+1)}\right\}\right)^{-1}. \quad (3.14)$$

In the same way, the maximum is attained for λ_k

$$\lambda_k^{(t+1)} = N\left(\sum_{i=1}^{N}\left(\left(\mathbf{m}_k^{(t+1)}(i)\right)^2 + \mathbf{C}_k^{(t+1)}(i,i)\right)\langle a_k(i)\rangle_{q^{(t+1)}(a_k(i))}\right)^{-1}. \quad (3.15)$$

Finally, taking the derivative with respect to ν_k and equating to zero, we find the "degrees of freedom" parameter of the Student-t by solving the equation

$$\frac{1}{N}\left(\sum_{i=1}^{N}\log\langle a_k(i)\rangle_{q^{(t+1)}(\tilde{\mathbf{a}})} - \sum_{i=1}^{N}\langle a_k(i)\rangle_{q^{(t+1)}(\tilde{\mathbf{a}})}\right) + \psi\left(\nu_k^{(t)}\frac{1}{2} + \frac{1}{2}\right) - \log\left(\nu_k^{(t)}\frac{1}{2} + \frac{1}{2}\right) - \psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1 = 0 \quad (3.16)$$

for ν_k, where

$$\psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

is the digamma function and $\nu_k^{(t)}$ is the value of ν_k at the previous iteration (t) used to evaluate the expectations in (3.13) during the VE-step.

IV. Computational Implementation

In our implementation, the variance of the additive noise is estimated in a preprocessing step and is kept fixed. The EM algorithm with a stationary Gaussian prior [3] and one output (the Laplacian operator) was used for this purpose. Furthermore, the EM-restored image was used to initialize our algorithm. For all experiments, four filter outputs $P = 4$ were used for the prior. We show the magnitude of the frequency responses of these filters in Fig. 2. The operators $\mathbf{Q}_1$ and $\mathbf{Q}_2$ correspond to the horizontal and vertical first order differences. Thus, these filters are used to model the vertical and horizontal image edge structure, respectively. The other two operators $\mathbf{Q}_3$ and $\mathbf{Q}_4$ are used to model the diagonal edge component contained in the vertical and horizontal edges, respectively. These filters are obtained by convolving the previous horizontal and vertical first order differences filters with fan filters with vertical and horizontal pass-bands, respectively. In our experiments, the fan filters in [26] were used.

We solve (3.10) and (3.16) iteratively. For (3.16), we employ the bisection method, as also proposed in [27]. In the next few paragraphs, we analyze how (3.10) is solved by a method based on the Lanczos process [29], [30].

Omitting the subscripts k and superscripts t for convenience, we regard (3.9) as the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A} = \mathbf{R}^{-1}$ is symmetric and positive definite, $\mathbf{b} = \beta\mathbf{H}^{\mathrm{T}}\mathbf{g}, \mathbf{x} = \mathbf{m}$, and products $\mathbf{A}\mathbf{v}$ can be obtained efficiently for any given $\mathbf{v}$. In addition, we have the linear algebra problem of estimating the diagonals of matrix $\mathbf{C} = \mathbf{Q}\mathbf{A}^{-1}\mathbf{Q}^{\mathrm{T}}$ in (3.13). The matrix $\mathbf{A} = \mathbf{R}^{-1}$ is very large; for example for 256×256 images it is of dimension $N \times N$ with $N = 65\,536$ and clearly an iterative method must be used.

The Lanczos process is an iterative procedure for transforming $\mathbf{A}$ to tridiagonal form [32]. Given some starting vector $\mathbf{b}$, it generates vectors $\{\boldsymbol{v}_n\}$ and scalars $\{\alpha_n, \beta_n\}$ as follows.

1. Set $\beta_1\boldsymbol{v}_1 = \mathbf{b}$ (meaning $\beta_1 = \|\mathbf{b}\|_2$ and $\boldsymbol{v}_1 = \mathbf{b}/\beta_1$ but exit if $\beta_1 = 0$).
2. For $n = 1, 2, \ldots$, set $\mathbf{w} = \mathbf{A}\boldsymbol{v}_n, \alpha_n = \boldsymbol{v}_n^T\mathbf{w}$, $\beta_{n+1}\boldsymbol{v}_{n+1} = w - a_n\boldsymbol{v}_n - \beta_n\boldsymbol{v}_{n-1}$.

After n steps, the situation can be summarized as

$$\mathbf{A}\mathbf{V}_n = \mathbf{V}_n\mathbf{T}_n + \beta_{n+1}\boldsymbol{v}_{n+1}\mathbf{e}_n^T \quad (4.1)$$

$$\mathbf{V}_n = [\,\boldsymbol{v}_1 \;\; \boldsymbol{v}_2 \;\; \cdots \;\; \boldsymbol{v}_n\,]$$

$$\mathbf{T}_n = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \beta_3 & \\ & \ddots & \ddots & \ddots \\ & & & \alpha_n \end{pmatrix} \quad (4.2)$$

where $\mathbf{e}_n$ is the nth unit vector, $\mathbf{V}_n$ has theoretically orthonormal columns, and $\mathbf{T}_n$ is tridiagonal and symmetric. In practice, $\mathbf{V}_n^T\mathbf{V}_n \neq \mathbf{I}$ unless $\boldsymbol{v}_{n+1}$ is reorthogonalized with respect to previous vectors, but relation (4.1) remains accurate to machine precision. This permits $\mathbf{V}_n$ and $\mathbf{T}_n$ to be used to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ accurately in a manner that is algebraically equivalent to the conjugate-gradients method, as described in [30] (It also leads to reliable methods for solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ when $\mathbf{A}$ is indefinite [30]). Note that $\boldsymbol{v}_1$ must be proportional to $\mathbf{b}$ as shown.

TABLE I
ISNR RESULTS COMPARING THE PROPOSED ALGORITHM WITH THE ALGORITHMS IN [9], [10], AND [11] USING THREE IMAGES, THREE NOISE LEVELS, AND GAUSSIAN SHAPED BLUR. THE ISNR RESULTS FOR THE BFO1, BFO2, BMK1, AND BMK2 ALGORITHMS ARE OBTAINED FROM [11]

Gaussian shaped blur with $\sigma^2 = 9$		Lena	Cameraman	Shepp-Logan
BSNR	Method	ISNR	ISNR	ISNR
40dB	*Stationary*	*3.33*	*2.44*	*3.56*
	Proposed	***4.86***	*3.45*	***9.46***
	TV-TE	*4.87*	*3.34*	*8.30*
	BFO1	*4.72*	***3.51***	*7.07*
	BFO2	*4.50*	*3.27*	*5.88*
	BMK1	*4.78*	*3.39*	*6.69*
	BMK2	*4.49*	*3.26*	*5.63*
30dB	*Stationary*	*2.54*	*1.89*	*2.80*
	Proposed	***3.89***	*2.74*	***5.94***
	TV-TE	*3.82*	*2.82*	*5.50*
	BFO1	*3.87*	***2.89***	*5.15*
	BFO2	*3.56*	*2.47*	*3.94*
	BMK1	*3.87*	*2.63*	*4.31*
	BMK2	*3.55*	*2.41*	*3.72*
20dB	*Stationary*	*2.23*	*1.43*	*2.14*
	Proposed	*2.76*	*1.86*	***3.92***
	TV-TE	*3.20*	*2.27*	*3.75*
	BFO1	***3.02***	*2.13*	*3.56*
	BFO2	*2.47*	***2.23***	*2.20*
	BMK1	*2.87*	*1.72*	*1.85*
	BMK2	*2.42*	*1.42*	*2.05*

When $\mathbf{A}$ is positive definite, each $\mathbf{T}_n$ is also positive definite and we may form the Cholesky factorization $\mathbf{T}_n = \mathbf{L}_n\mathbf{L}_n^T$ (with $\mathbf{L}_n$ lower-triangular) by updating $\mathbf{L}_{n-1}$. The conjugate-gradient method computes a sequence of approximate solutions to $\mathbf{Ax} = \mathbf{b}$ in the form $\mathbf{x}_n = \mathbf{V}_n\mathbf{y}_n$, where $\mathbf{y}_n$ is defined by the equation $\mathbf{T}_n\mathbf{y}_n = \beta_1\mathbf{e}_1$. Since $\mathbf{V}_n(\beta_1\mathbf{e}_1) = \mathbf{b}$ exactly for all n, we see from (4.1) that $\mathbf{Ax}_n = \mathbf{b} + \mathbf{r}_n$, where the residual vector $\mathbf{r}_n = \beta_{n+1}\boldsymbol{\upsilon}_{n+1}(\mathbf{e}_n^T\mathbf{y}_n)$ becomes small if either β_{n+1} is small (unlikely in practice) or the last element of $\mathbf{y}_n$ is small.

In practice, we do not compute $\mathbf{y}_n$ itself because every element differs from $\mathbf{y}_{n-1}$. Instead, we compute two quantities $\mathbf{z}_n$ and $\mathbf{W}_n$ by applying forward substitution to the lower-triangular systems $\mathbf{L}_n\mathbf{z}_n = \beta_1\mathbf{e}_1$ and $\mathbf{L}_n\mathbf{W}_n^T = \mathbf{V}_n^T$, where

$$\mathbf{z}_n = \begin{bmatrix} \mathbf{z}_{n-1} \\ \zeta_n \end{bmatrix}, \quad \mathbf{W}_n = [\,\mathbf{W}_{n-1} \quad \mathbf{w}_n\,] \equiv \mathbf{V}_n\mathbf{L}_n^{-T} \tag{4.3}$$

so that $\mathbf{x}_n$ can be updated according to $\mathbf{x}_n = \mathbf{V}_n\mathbf{y}_n = \mathbf{W}_n\mathbf{z}_n = \mathbf{x}_{n-1} + \zeta_n\mathbf{w}_n$. Since $\mathbf{L}_n$ is bidiagonal, only the most recent columns of $\mathbf{V}_n$ need to be retained in memory. Thus, the previous equation is the update rule for the image estimate in the algorithm.

In order to estimate elements of $\mathbf{A}^{-1}$, we can make use of the same vectors $\mathbf{w}_n$ in (4.3). If *we now assume* that exact arithmetic holds, we see that

$$\begin{aligned}\mathbf{W}_n^T\mathbf{A}\mathbf{W}_n &= \mathbf{L}_n^{-1}\mathbf{V}_n^T\mathbf{A}\mathbf{V}_n\mathbf{L}_n^{-T} \\ &= \mathbf{L}_n^{-1}\mathbf{T}_n\mathbf{L}_n^{-T} = \mathbf{L}_n^{-1}\mathbf{L}_n\mathbf{L}_n^T\mathbf{L}_n^{-T} = \mathbf{I}.\end{aligned}$$

If *we further assume* that the Lanczos process continues for N iterations, we have $\mathbf{W}_N^T\mathbf{A}\mathbf{W}_N = \mathbf{I}$, so that $\mathbf{W}_N\mathbf{W}_N^T = \mathbf{A}^{-1}$. On this basis, if we define $\mathbf{B}_n = \mathbf{W}_n\mathbf{W}_n^T$, we have the sequence of estimates $\mathbf{B}_n = \mathbf{B}_{n-1} + \mathbf{w}_n\mathbf{w}_n^T \approx \mathbf{A}^{-1}$. To estimate its ith diagonal, we form the sum $\mathbf{e}_i^T\mathbf{B}_n\mathbf{e}_i = \sum_n \mathbf{w}_{in}^2$.

TABLE II
ISNR RESULTS COMPARING THE PROPOSED ALGORITHM WITH THE ALGORITHMS IN [9], [10], AND [11] USING THREE IMAGES, THREE NOISE LEVELS, AND UNIFORM BLUR. THE ISNR RESULTS FOR THE BFO1, BFO2, BMK1, AND BMK2 ALGORITHMS ARE OBTAINED FROM [11]

Uniform 9×9 blur		Lena	Cameraman	Shepp-Logan
BSNR	Method	ISNR	ISNR	ISNR
40dB	*Stationary*	*4.72*	*4.57*	*5.31*
	Proposed	***8.49***	***9.53***	***15.08***
	TV-TE	*8.43*	*9.07*	*16.63*
	BFO1	*8.34*	*8.55*	*14.22*
	BFO2	*8.35*	*8.25*	*12.01*
	BMK1	*8.42*	*8.57*	*13.69*
	BMK2	*8.37*	*8.46*	*12.05*
30dB	*Stationary*	*4.06*	*3.24*	*3.56*
	Proposed	***6.10***	***6.29***	***9.71***
	TV-TE	*5.93*	*6.26*	*10.66*
	BFO1	*6.08*	*5.68*	*8.88*
	BFO2	*5.64*	*4.65*	*6.91*
	BMK1	*5.89*	*5.41*	*7.77*
	BMK2	*5.58*	*4.38*	*6.50*
20dB	*Stationary*	*2.68*	*2.19*	*2.49*
	Proposed	*3.98*	***3.33***	***6.10***
	TV-TE	*3.90*	*3.33*	*6.26*
	BFO1	*4.09*	*3.31*	*5.57*
	BFO2	***4.14***	*2.12*	*2.95*
	BMK1	*3.72*	*2.42*	*3.01*
	BMK2	*3.15*	*1.94*	*2.64*

Thus, we can obtain monotonically increasing estimates for all diagonals at very little cost,[1] in the manner of LSQR [33].

Similarly, for the matrix $\mathbf{C}$, whose diagonals we wish to estimate, we have

$$\begin{aligned}\mathbf{e}_i^T\mathbf{C}\mathbf{e}_i &= \mathbf{e}_i^T\mathbf{Q}\mathbf{A}^{-1}\mathbf{Q}^T\mathbf{e}_i \\ &\approx \mathbf{e}_i^T\mathbf{Q}\sum_n\left(\mathbf{w}_n\mathbf{w}_n^T\right)\mathbf{Q}^T\mathbf{e}_i \\ &= \mathbf{e}_i^T\sum_n\left(\mathbf{q}_n\mathbf{q}_n^T\right)\mathbf{e}_i = \sum_n \mathbf{q}_{in}^2\end{aligned}$$

where $\mathbf{q}_n = \mathbf{Q}\mathbf{w}_n$ can be formed at each Lanczos iteration and then discarded after use. This is how we evaluate $\mathbf{C}_k^{(t)}(i,i)$ in (3.12).

Element estimation of inverses of large matrices is also required in many other recently developed Bayesian algorithms (see for example [11], [19], and [23]) and presently to the best of our knowledge are handled either by inaccurate circulant or diagonal approximations of the matrix $\mathbf{A}$ or by very time-consuming Monte-Carlo approaches.

An iteration of the variational EM algorithm consists of the update steps given by (3.9)–(3.12) and (3.14)–(3.15). In our implementation, the parameter β is estimated in a preprocessing step, as described above. During the variational M-step the bisection method is used for the update of the parameters ν_k with termination criterion $|\nu_k^m - \nu_k^{m-1}| < 10^{-6}$, where ν_k^m is the value of ν_k at the mth iteration of the bisection method. The linear system in (3.10) is solved by the iterative Lanczos procedure. The termination criterion for this algorithm is

$$\begin{aligned}\left\|\mathbf{r}_n^{(t)}\right\| &= \left\|\mathbf{H}^{\mathrm{T}}\mathbf{g} - \mathbf{R}^{(t-1)}\mathbf{m}_n^{(t)}\right\| \\ &< \left\|\left(\mathbf{R}^{(t-1)}\right)^{-1}\right\|_{\mathrm{fro}}\left\|\mathbf{m}_n^{(t)}\right\| 10^{-9}\end{aligned}$$

[1]See http://www.stanford.edu/group/SOL/software/cgLanczos.html for Matlab code.

Fig. 1. (a) Degraded "Cameraman" image by uniform 9×9 blur and noise with $\mathrm{BSNR} = 40$ dB, (b) restored image using a stationary Gaussian prior [3] $\mathrm{ISNR} = 4.57$ dB, (c) restored image using $\mathrm{TV-TE}$ $\mathrm{ISNR} = 9.07$ dB, (d) restored image using proposed algorithm $\mathrm{ISNR} = 9.53$ dB.

where n denotes the iteration index of the Lanczos process (hence, $n = 1, 2, \ldots, M^{(t)}$). Thus, $\mathbf{m}_n^{(t)}$ is the image estimate at the n-th Lanczos iteration and at the t-th iteration of the overall variational algorithm. Lastly, $||\cdot||_{\mathrm{fro}}$ denotes the *Frobenius* norm. As criterion for termination of the variational algorithm we used $||\mathbf{r}_{M^{(t+1)}}^{(t+1)}|| \geq ||\mathbf{r}_{M^{(t)}}^{(t)}||$. In other words, we terminate the overall algorithm when the residual of the Lanczos process at iteration $t+1$ is larger than that of the iteration t.

The overall algorithm is summarized in the following three-step procedure.

1. Initialize $\mathbf{m}^0, \beta$ using a stationary model [3].
2. Repeat until convergence:
   t-th iteration:
   - VE-step: Update, $\mathbf{m}^t, \mathbf{R}^t$ and $q^t(\mathbf{a}_k)$ using (3.10), (3.12) and (3.13), respectively. For the last equation, $\mathbf{m}_k^{(t)}$ and $\mathbf{C}_k^{(t)}$ also need to be calculated. Also, calculate the expected value of $q^t(a_k(i))$ from (3.13), need for the VM-step and the next VE-step in the $(t+1)$th iteration.
   - VM-step: Update λ_k^t using (3.15) and ν_k^t by solving (3.16) for each k.
3. Use $\mathbf{m}^t$ as the restored image estimate.

V. Numerical Experiments

We demonstrate the value of the proposed restoration approach by showing results from various experiments with three 256×256 input images: "Lena," "Cameraman," and "Shepp–Logan" phantom. Every image is blurred with two types of blur; the first has the shape of a Gaussian function with shape parameter 9, and the second is uniform with support a rectangular region of 9×9 pixels. The blurred signal to noise ratio (BSNR) defined as follows was used to quantify the noise level:

$$\mathrm{BSNR} = 10 \log_{10} \frac{||\mathbf{Hf}||^2}{\sigma^2 N}$$

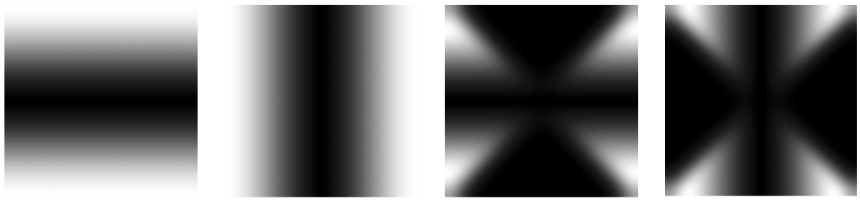
where σ^2 is the variance of the additive white Gaussian noise (AWGN). Three levels of AWGN were added to the blurred images with $\mathrm{BSNR} = 40$, 30, and 20 dB. Thus, in total 18 image restoration experiments were performed to test the proposed algorithm.

As performance metric, the improvement in signal-to-noise ratio (ISNR) was used

$$\mathrm{ISNR} = 20 \log_{10} \frac{||\mathbf{f} - \mathbf{g}||}{||\mathbf{f} - \hat{\mathbf{f}}||}$$

where $\mathbf{f}, \mathbf{g}$ and $\hat{\mathbf{f}}$ are the original, observed degraded and restored images, respectively.

We present ISNR results comparing our algorithm with four total-variation (TV) based Bayesian algorithms in [10] abbreviated as BFO1, in [9] abbreviated as BFO2, and [11] abbreviated



(a) (b) (c) (d)

Fig. 2. Magnitude of frequency responses of the filters used in the prior: (a) horizontal differences ($\mathbf{Q}_1$), (b) vertical differences ($\mathbf{Q}_2$), (c) $\mathbf{Q}_3$ and (d) $\mathbf{Q}_4$.

TABLE III
ISNR RESULTS COMPARING THE PROPOSED ALGORITHM WITH THE ALGORITHMS IN [9] USING 3 IMAGES, 3 NOISE LEVELS AND BLUR $[1\ 4\ 6\ 4\ 1]^{\mathrm{T}}[1\ 4\ 6\ 4\ 1]/256$

Pyramidal blur		Lena	Cameraman	Shepp-Logan
BSNR	Method	*ISNR*	*ISNR*	*ISNR*
40dB	*Stationary*	*4.82*	*3.82*	*2.68*
	Proposed	***7.02***	***6.40***	***13.70***
	BFO1	*5.56*	*6.07*	*10.87*
30dB	*Stationary*	*3.03*	*2.45*	*1.55*
	Proposed	***4.81***	*4.25*	***8.51***
	BFO1	*4.52*	***4.35***	*7.91*
20dB	*Stationary*	*1.57*	*1.26*	*1.01*
	Proposed	***3.03***	***2.75***	***7.00***
	BFO1	*3.01*	*2.60*	*5.91*

as BMK1 and BMK2. For comparison purposes we also implemented a restoration algorithm based on TV regularization [8]. This algorithm minimizes the function $J(\mathbf{f})$ with respect to the image

$$J(\mathbf{f}) = \|\mathbf{g} - \mathbf{H}\mathbf{f}\|^2 + \lambda \sum_{k=1}^{N} \sqrt{(D_h\mathbf{f})_k^2 + (D_v\mathbf{f})_k^2 + c}$$

where $D_x\mathbf{f}$ and $D_y\mathbf{f}$ are the directional differences vectors of the image along the horizontal and vertical direction respectively. A conjugate gradient algorithm is used to minimize $J(\mathbf{f})$ with a one-step-late quadratic approximation [8]. The parameters λ and c were kept fixed during the iterations of this algorithm and were selected by trial-and-error (TE) to optimize ISNR assuming knowledge of the original image. Since this algorithm assumes knowledge of the original it is not a realistic one. However, it provides the performance bound of the TV algorithm with fixed parameters. In Tables I and II, we present ISNR results comparing our algorithm with the above-mentioned methods in 18 experiments. The ISNR results for BFO1, BFO2, BMK1, and BMK2 were obtained from [11]. In these tables for reference purposes we also provide ISNR results for the stationary simultaneously autoregressive prior in [3].

In Fig. 1, restoration results are shown for the "Cameraman" image with $\mathrm{BSNR} = 40$ dB noise and uniform blur. In this experiment the restored image by the proposed algorithm is superior in ISNR, and is visually distinguishable from the TV-TE approach, which was optimized using the original image.

At this point we note that the proposed algorithm performed very well compared with the TV-based methods in [9], [10], and [11]. More specifically, for the high $\mathrm{BSNR} = 40$ dB case it gave the best results from all methods (excluding TV-TE since it is unrealistic) in 5 out of 6 experiments. For the midlevel $\mathrm{BSNR} = 30$ dB case it gave the best performance in 5 out of 6 experiments. Finally, in the low $\mathrm{BSNR} = 20$ dB case it gave the best result in 3 out of the 6 experiments. Overall the proposed algorithm gave the best ISNR results in 13 out of 18 experiments, compared to 3 out of 18 for BFO1 and 2 out of 18 for BFO2.

We also compared our method with BFO1 [9], which based on the above experiments was the most competitive TV based method. We used the same three images and noise levels as above. We also used a 5×5 pyramidal blur with impulse response given by $[1\ 4\ 6\ 4\ 1]^T[1\ 4\ 6\ 4\ 1]/256$. The ISNR results for this experiment are given in Table III. For the implementation we used the code provided by the authors.[2] The ISNR results from this experiment are consistent with the previous ones.

VI. CONCLUSIONS AND FUTURE RESEARCH

We presented a new Bayesian framework for image restoration that uses a product-based Student-t type of priors. The main theoretical contribution is that by constraining the approximation of the posterior in the variational framework, we bypass the need for knowing the normalization constant of this

[2]http://www.lx.it.pt/~jpaos

prior. Thus, we avoid having to use improper priors, i.e., priors whose normalization constant is empirically selected; see, for example, [9]–[11], [17], [18], and [19]. Furthermore, the proposed methodology does not require empirical parameter selection as in the MAP methodology that uses a similar-in-spirit prior in [17] and [18]. We also presented a Lanczos-based computational scheme tailored to the computations required by our algorithm.

We demonstrated by the ISNR results in Tables I–III that the proposed method is competitive with the very recently proposed TV-based Bayesian algorithms in [9], [10], and [11]. More specifically, it appears that this approach is more competitive in the higher BSNR cases. Thus, it seems that in such cases the proposed Student-t model has the ability to capture more accurately than TV-based priors subtle features of the image present in the observations. However, in the presence of high levels of AWGN this does not seem to be the case and the advantage of our proposed prior compared to TV priors seems to diminish. We believe that this is the case because high levels of noise "wipe out" the subtle features that our model can capture.

We found empirically that modeling explicitly the diagonal edge structure contained in the vertical and horizontal edge (the use of operators $\mathbf{Q}_3$ and $\mathbf{Q}_4$) improved the performance of the proposed algorithm, for a wide range of images, blurs and SNRs. Selecting optimally such operators according to the image is a topic of current investigation.

Another topic of current investigation is image models that capture the spatial correlation between the outputs of the convolutional filters used in the prior. We plan to address this point by assuming a similar-in-spirit prior that uses a neighborhood around each pixel and multidimensional Student-t pdfs. Another point that we plan to investigate is the use of *generalized* Student-t pdfs. These pdfs depend on $|x|^c$ and the "classical" Student-t used herein is just a special case with $c = 2$.

APPENDIX

In the VE-step the bound must be optimized with respect to $\mathbf{R}, \mathbf{m}$ and $q(\tilde{\mathbf{a}})$. With the mean field approximation (3.6) the bound becomes

$$\begin{aligned} &L(q(\tilde{\mathbf{a}}), \theta_1, \theta_2) \\ &= \int \prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k) \log p(\tilde{\mathbf{y}}, \tilde{\varepsilon}, \tilde{\mathbf{a}}; \theta_2) d\tilde{\varepsilon} d\tilde{\mathbf{a}} \\ &\quad - \int \prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k) \log \prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k) d\tilde{\varepsilon} d\tilde{\mathbf{a}} \end{aligned}$$

where $\theta_1 = [\mathbf{R}, \mathbf{m}]$ and $\theta_2 = [\beta, \lambda_1, \ldots, \lambda_P, \nu_1, \ldots, \nu_P]^{\mathrm{T}}$.

Because at this point we aim to optimize with respect to θ_1, we operate on the function L', which includes only the terms that depend on the parameters θ_1

$$\begin{aligned} L &\propto L'(\theta_1) \\ L'(\theta_1) &= \sum_{k=1}^{P} \int q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k) \\ &\quad \times \log p(\mathbf{y}_k|\boldsymbol{\varepsilon}_k;\theta_2) p(\boldsymbol{\varepsilon}_k|\mathbf{a}_k;\theta_2) d\boldsymbol{\varepsilon}_k d\mathbf{a}_k \\ &\quad - \sum_{k=1}^{P} \int q(\boldsymbol{\varepsilon}_k;\theta_1) \log q(\boldsymbol{\varepsilon}_k;\theta_1) d\boldsymbol{\varepsilon}_k. \end{aligned} \tag{A.1}$$

The first sum is further analyzed

$$\begin{aligned} &\sum_{k=1}^{P} \int q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k) \log p(\mathbf{y}_k|\boldsymbol{\varepsilon}_k;\theta_2) \\ &\quad \times p(\boldsymbol{\varepsilon}_k|\mathbf{a}_k;\theta_2) d\boldsymbol{\varepsilon}_k d\mathbf{a}_k \\ &\quad \propto \sum_{k=1}^{P} \big\langle -\beta(\mathbf{H}\boldsymbol{\varepsilon}_k - \mathbf{y}_k)^{\mathrm{T}} \mathbf{Q}_k^{-\mathrm{T}} \mathbf{Q}_k^{-1} \\ &\quad \times (\mathbf{H}\boldsymbol{\varepsilon}_k - \mathbf{y}_k) - \lambda_k \boldsymbol{\varepsilon}_k^{\mathrm{T}} \mathbf{A}_k \boldsymbol{\varepsilon}_k \big\rangle_{q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k)} \\ &\quad = -\beta P \|\mathbf{H}\mathbf{m} - \mathbf{g}\|_2^2 - \sum_{k=1}^{P} \lambda_k \mathbf{m}^{\mathrm{T}} \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k \mathbf{Q}_k \mathbf{m} \\ &\quad - \mathrm{trace}\left\{ \left(\beta P \mathbf{H}^{\mathrm{T}}\mathbf{H} + \sum_{k=1}^{P} \lambda_k \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k \mathbf{Q}_k \right) \mathbf{R} \right\} \end{aligned} \tag{A.2}$$

where $\hat{\mathbf{A}}_k$ is a diagonal matrix with elements

$$\hat{\mathbf{A}}_k(i,i) = \langle a_k(i) \rangle_{q(a_k(i))}, i = 1, \ldots, N.$$

The second integral is the entropy of a Gaussian function, which is proportional to

$$\int q(\boldsymbol{\varepsilon}_k;\theta_1) \log q(\boldsymbol{\varepsilon}_k;\theta_1) d\boldsymbol{\varepsilon}_k \propto \frac{1}{2} \log\det |\mathbf{R}|. \tag{A.3}$$

Setting the derivative of L' w.r.t $\mathbf{R}$ equal to zero using (A.1)–(A.3) yields the equation shown at the bottom of the page.

Similarly, using (A.2), we find that the optimum for the mean

$$\frac{\partial L'(\theta_1)}{\partial \mathbf{m}} = 0 \Rightarrow \mathbf{m} = \beta \mathbf{R} \mathbf{H}^{\mathrm{T}} \mathbf{g}.$$

The final part of the VE-step is the optimization w.r.t. the function $q(\tilde{\mathbf{a}})$. It is straightforward to verify that this is achieved

$$\begin{aligned} \frac{\partial L'(\theta_1)}{\partial \mathbf{R}} = 0 \Rightarrow \frac{\partial \mathrm{trace}\left\{ \beta P \mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{R} + \sum_{k=1}^{P} \lambda_k \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k \mathbf{Q}_k \mathbf{R} \right\} - P \partial \log\det |\mathbf{R}|}{\partial \mathbf{R}} = 0 \\ \Rightarrow \beta P \mathbf{H}^{\mathrm{T}}\mathbf{H} + \sum_{k=1}^{P} \lambda_k \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k \mathbf{Q}_k - P\mathbf{R}^{-1} = 0 \Rightarrow \mathbf{R} = \left(\beta \mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{1}{P} \sum_{k=1}^{P} \lambda_k \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k \mathbf{Q}_k \right)^{-1} \end{aligned}$$

when

$$q(\tilde{\mathbf{a}}) = \frac{\exp(\langle \log p(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}) \rangle_{q(\tilde{\boldsymbol{\varepsilon}})})}{\int \exp(\langle \log p(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}) \rangle_{q(\tilde{\boldsymbol{\varepsilon}})}) d(\tilde{\mathbf{a}})} = \prod_{k=1}^{P} \prod_{i=1}^{N} q(a_k(i)).$$

The product form is due to

$$\exp\langle \log p(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}) \rangle_{q(\tilde{\boldsymbol{\varepsilon}})} \propto \prod_{k=1}^{P} \prod_{i=1}^{N} (a_k(i))^{\frac{\nu_k}{2}+\frac{1}{2}-1} \exp\left\{ -\frac{\nu_k}{2} a_k(i) - \frac{1}{2}\lambda_k((\mathbf{m}_k(i))^2 + \mathbf{C}_k(i,i)) a_k(i) \right\}.$$

Hence, each $q(a_k(i))$ is a Gamma distribution

$$q(a_k(i)) = \text{Gamma}\left(a_k(i); \frac{\nu_k}{2} + \frac{1}{2}, \frac{\nu_k}{2} + \frac{1}{2}\lambda_k \times ((\mathbf{m}_k(i))^2 + \mathbf{C}_k(i,i)) \right)$$

where $\mathbf{m}_k = \mathbf{Q}_k \mathbf{m}$ and $\mathbf{C}_k = \mathbf{Q}_k \mathbf{R} \mathbf{Q}_k^{\mathrm{T}}$.

REFERENCES

[1] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structures and problems," *IEEE Trans. Signal Process.*, vol. 37, no. 12, pp. 2024–2036, Dec. 1989.

[2] N. P. Galatsanos and A. K. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Trans. Image Process.*, vol. 1, no. 3, pp. 322–336, Jul. 1992.

[3] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyper-parameter estimation in image restoration," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 231–246, Feb. 1999.

[4] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge preserving MAP estimation," *IEEE Trans. Image Process.*, vol. 2, no. 3, pp. 296–310, Jul. 1993.

[5] R. R. Schultz and R. L. Stevenson, "A Bayesian approach to image expansion with improved resolution," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 233–242, May 1994.

[6] R. Pan and S. Reeves, "Efficient Huberov edge preserving image restoration," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3728–3735, Dec. 2006.

[7] J. Portilla, V. Strella, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[8] T. F. Chan, S. Esedoglu, F. Park, and M. H. Yip, "Recent developments in total variation image restoration," in *Handbook of Mathematical Models in Computer Vision*. New York: Springer, 2005.

[9] J. Bioucas-Dias, M. Figueiredo, and J. Oliveira, "Adaptive Bayesian/total-variation image deconvolution: A majorization-minimization approach," presented at the Eur. Signal Processing Conf.—EUSIPCO, Florence, Italy, Sep. 2006.

[10] J. Bioucas-Dias, M. Figueiredo, and J. Oliveira, "Total-variation image deconvolution: A majorization-minimization approach," presented at the Int. Conf. Acoustics and Speech and Signal Processing, ICASSP, May 2006.

[11] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Parameter estimation in TV image restoration using variational distribution approximation," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 326–339, Mar. 2008.

[12] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 866–881, Aug. 2003.

[13] J. M. Bioucas-Dias, "Bayesian wavelet-based image deconvolution: A GEM algorithm exploiting a class of heavy-tailed priors," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 937–951, Apr. 2006.

[14] K. Hirakawa and X.-L. Meng, "An empirical bayes EM-wavelet unification for simultaneous denoising, interpolation, and/or demosaicing," presented at the IEEE Int. Conf. Image Processing, Atlanta, GA, Sep. 2006.

[15] S. Roth and M. J. Black, "Fields of experts: A framework for learning image priors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2005, vol. II, pp. 860–867.

[16] D. Sun and W.-K. Cham, "Postprocessing of low bit-rate block DCT coded images based on a fields of experts prior," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2743–2751, Nov. 2007.

[17] G. Chantas, N. P. Galatsanos, and A. Likas, "Bayesian restoration using a new nonstationary edge-preserving image prior," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 2987–2997, Oct. 2006.

[18] G. K. Chantas, N. P. Galatsanos, and N. Woods, "A super-resolution based on fast registration and maximum *a posteriori* reconstruction," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1821–1830, Jul. 2007.

[19] D. Tzikas, A. Likas, and N. Galatsanos, "Variational Bayesian blind image deconvolution with student-t priors," presented at the IEEE Int. Conf. Image Processing, San Antonio, TX, Sep. 2007.

[20] A. Kanemura, S.-I. Maeda, and S. Ishii, "Hyperparameter estimation in Bayesian image superresolution with a compound Markov random field prior," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing*, Thessaloniki, Greece, Aug. 2007, pp. 181–186.

[21] A. Mairgiotis, N. Galatsanos, and Y. Yang, "New detectors for watermarks with unknown power based on student-t image priors," presented at the IEEE Int. Conf. Multimedia Signal Processing, MMSP, Chania, Crete, 2007.

[22] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer Verlag, 2006.

[23] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[24] C. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications.*. New York: Wiley, 1995.

[25] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback–Leibler distance," *IEEE Trans. Image Process.*, vol. 11, no. 2, pp. 146–158, Feb. 2002.

[26] A. L. Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3089–3101, Oct. 2006.

[27] C. Liu and D. B. Rubin, "ML estimation of the t distribution using EM and its extensions," *ECM and ECME, Statist. Sin.*, vol. 5, pp. 19–39, 1995.

[28] A. Likas and N. Galatsanos, "A variational approach for Bayesian blind image deconvolution," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2222–2233, Aug. 2004.

[29] M. Beal, "Variational Algorithms for Approximate Bayesian Inference," Ph.D. Dissertation, The Gatsby Computational Neuroscience Unit, University College, London, U.K., 2003.

[30] C. C. Paige and M. A. Saunders, "Solution of sparse indefinite systems of linear equations," *SIAM J. Numer. Anal.*, vol. 12, pp. 617–629, 1975.

[31] Y. Saad, *Iterative Methods for Sparse Linear Systems, Second Edition*. Philadelphia, PA: SIAM, 2000.

[32] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[33] C. C. Paige and M. A. Saunders, "LSQR: An algorithm for sparse linear equations and sparse least squares," *ACM Trans. Math. Softw.*, vol. 8, no. 1, pp. 43–71, 1982.

Giannis Chantas, photograph and biography not available at the time of publication.

Nikolaos Galatsanos, photograph and biography not available at the time of publication.

Aristidis Likas, photograph and biography not available at the time of publication.

Michael Saunders, photograph and biography not available at the time of publication.